Resampling Methods to Handle the Class-Imbalance Problems in Predicting Protein-Protein Interaction Site and Beta-Turn

(タンパク質間相互作用予測およびβターン予測におけるクラス不均衡問題を扱うためのリサンプリング手法)

NGUYEN THI LAN ANH

Graduate School of Natural Science and Technology
Kanazawa University

Abstract

Proteins are the active functional biomolecules that are responsible for many tasks in the cells. Most proteins interact with the other proteins or molecules to perform their functions.

Though many protein sequences have not been determined, the functions of the unknown protein can be inferred from the functions of the known proteins that interact with it. In addition, functions of a protein directly depend on its three-dimensional structure. Therefore, studying of protein-protein interaction (PPI) and protein structure is very important in bioinformatics.

The study of PPI sites aims to localize where protein sequence can physically interact. Learning about this issue leads to the understanding how proteins recognize the other molecules.

Predicting β-turns and their types is one of the interesting and hard problems in bioinformatics, in order to provide more information for fold recognition study. However, the performances of both β-turns prediction and PPI sites prediction are still far from being perfect. One of the main reasons is the existence of class-imbalance problem in the datasets. This thesis intends to enhance the performances of predicting (i) the protein-protein interaction site; and (ii) the β-turn and beta-turn's types by mainly focusing on this problem.

The experimental results on the imbalanced PPI site dataset showed a significant improvement of our method in comparison with the other state-of-the-art methods.

We performed experiments on three benchmark datasets to evaluate the performance of our method for predicting the β-turns and their types. The results showed the substantial improvement of our approach compared with the other strategies.

Chapter 1 Introduction

1.1 Introduction

1.1.1 Protein overview

Protein

Proteins are cellular large molecules that are constructed from chains of hundreds or thousands amino acids. Each chain is called a polypeptide. Each individual amino acid is called a residue. Two amino acids link together through the peptide bond.

Proteins play a very important role in the cells of living organisms. Each protein has a specific function. Most proteins interact with the other molecules to perform their function. If the interactions between proteins in a cell disappear, the cell will be blind, deaf, paralytic and disintegrate.

Functions of proteins directly depend on their structure and shape. Protein structure can be presented as four levels a in Figure 1.1.

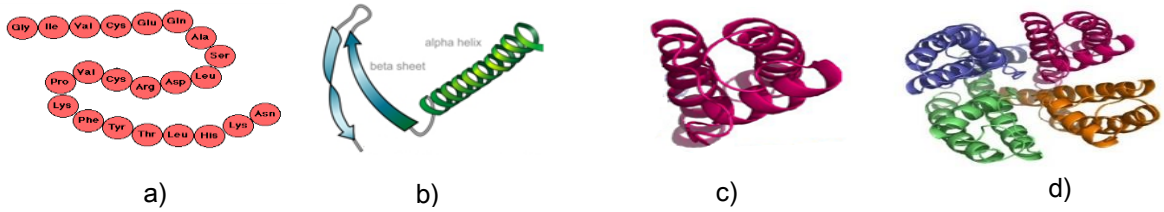

Figure 1.1 Four levels of protein structure.

a) Primary structure is a sequence of amino acids.
b) Secondary structure is the spatial arrangement of the specific regions.
c) Tertiary structure is the 3D structure of the whole polypeptide chain.
d) Quaternary structure, if exists, is the 3D structure of many polypeptide chains.

Protein blocks

Around 50% of total number protein residues are assigned as coils while they actually correspond to many distinct local protein structures. Therefore, a new view of three-dimensional protein structure that combines the small local fragments has been developed. A structural alphabet (SA) is a complete set of these prototypes.

Protein Blocks (PBs) [1] allows a good approximation of local protein 3D structures [2]. This SA is composed of sixteen local structure prototypes of five consecutive Cα, labelled a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, respectively. Each of these prototypes represents a vector of eight average dihedral angles ϕ/ψ.

1.1.2 Protein-protein interaction sites prediction

Protein-protein interactions are responsible for many important biological processes. The study of protein-protein interaction contains two main goals, recognizing the interaction sites, and predicting which pairs of proteins can interact. The knowledge of protein interfaces allows us to understand the way protein recognizes the other molecules and engineers new interactions. It is also very useful in identifying drug targets, designing drug-like peptides to prevent unwanted interactions.

Predicting protein-protein interaction sites by machine learning methods can be dealt as a classification problem that to predict whether an amino acid is an interface residue or not. The features that can distinguish interaction and non-interaction residues are used to describe protein sites.

Some studies have attempted to develop the techniques for predicting interaction sites from protein sequences. However, it is not easy to apply sequence-based methods for interaction sites prediction due to the lack of understanding of biological properties that can provide vital information related to binding sites. On the other hand, because the number of non-interacting residues is much more than the number of interacting residues, it often leads to the high value of false predicted negative.

1.1.3 β-turn prediction

A β-turn is composed of four consecutive residues that are not in an α-helix and the distance between the first and the fourth Cα is less than 7Å [3]. β-turns play an important role in the conformation as well as the function of protein, and make up around 25% of the residue numbers. Therefore, the knowledge of β-turn is very necessary in three-dimensional structure prediction of a given primary protein sequence.

β-turns are categorized into nine types (I, I', II, II', IV, VIa1, VIa2, VIb and VIII) based on the dihedral angles of residues i+1 and i+2 in the turn [4]. Because the turn types VIa1, VIa2 and VIb are rare, they are often combined into one type and named VI.

There are many methods for predicting β-turns and their types have been proposed. However, the quality of both β-turn location and turn types prediction is a challenge.

1.1.4 Class-imbalance problems

A dataset is imbalanced if the number of samples in some classes is significantly larger than in other classes. In the case of two-class datasets, the class with small amount of samples is the minority (positive) class while the other is the majority (negative) class.

Class-imbalance problem is very common in the field of bioinformatics. When applying standard machine learning to the such datasets, most of the learning systems can be seriously influenced and tend to predict majority class exactly while users desire for both high sensitivity and specificity.

Basically, the methods for handling class-imbalance problems are divided into two categories: data level methods, and algorithmic level methods. However, data level methods are said to be more effective on improving classifier accuracy than algorithm level methods.

1.2 Objectives

This thesis aims to (i) improve the performance of predicting protein interface residue by solving the problem of class-imbalance and using a new kind of feature for well distinguishing the protein interface and non-interface residues; (ii) to better the quality of predicting β-turns and their types by utilizing random under-sampling method to balance the dataset.

1.3 Contributions

The main contributions of this thesis are described as below:

Firstly, a novel over-sampling technique for relaxing the class-imbalance problem based on local density distributions**,** OSD, was proposed and applied for predicting protein-protein interaction site. We also proposed the methods combined with KSVM-THR and random under-sampling methods to reinforce the tolerance for the class imbalance problem. In addition, we found that the information of predicted shape strings increased the performance for predicting whether interface or non-interface residues.

Secondly, we utilize predicted protein blocks and position specific scoring matrix together with random under-sampling method to improve the prediction of the β-turns and their types compared with the state-of-the-art methods.

Chapter 2 Methods for Dealing with Class-imbalance Problems

2.1 Standard Classifier Modeling Algorithm

There are many basic well-known classifier learning algorithms. However, because of the limitation of space, we just focus on Support Vector Machines that are mainly used for our research.

Support Vector Machines (SVMs) [5] were proposed by Vapnik. SVMs originally were for the linear binary classification problem. However, in many applications, the linear classifier cannot work well but the non-linear classifier. In these cases, the non-linear separated problem is transformed into a high dimensional feature space using a set of non-linear basis functions.

SVMs are believed to be less affected by the class imbalance problem than other classification learning algorithms since boundaries between classes are calculated based on the support vectors and

the class sizes may not affect the class boundary too much. However, some weaknesses of SVMs when applying to the imbalanced datasets were reported.

2.2 The State-of-the-art Solutions for Class-imbalance Problems

2.2.1 Resampling techniques

Over-sampling

Over-sampling method tries to balance the data set by increasing the number of minority class samples.

The simplest way is Random Over-sampling, which randomly chooses some minority samples, replicates them and then adds to the original dataset. However, this method can lead to the over-fitting.

SMOTE [6] and their improvements such as SMOTEBoost, Smote-RSB, Safe-Level-SMOTE, Borderline-SMOTE, and so on, were proposed to overcome the over-fitting by generating synthetic samples from the minority class instances.

The other over-sampling methods that need to pay attention to are the Cluster-based sampling algorithms. These methods are more flexible than the simple and synthetic sampling algorithms, and can be tailored to target very specific problems.

Under-sampling

Under-sampling method solves the class-imbalance problem by decreasing the number of majority class samples, therefore, decreases the cost of computation.

Random Under-sampling balances the original data set distribution by randomly eliminating some majority samples. However, this way may lead to lose a lot of important information of the majority class. EasyEnsemble, BalanceCascade [7] were proposed to overcome this limitation.

The other under-sampling methods that based on k-nearest neighbors are NearMiss-1, NearMiss-2, Near-Miss-3, and the "most distant" method [8].

Anand et al. [9] introduced an under-sampling method that also based on nearest neighbor and weighted SVM. For each minority class sample, its k closest majority class samples will be removed.

2.2.2 Algorithm level methods for handling imbalance

A popular way for dealing with the class-imbalance problem is to choose a proper inductive bias. For decision trees, approaches are adjusting the probabilistic estimate at the tree leaf or developing new pruning techniques.

For SVMs, the use of different penalty constants for different classes, using the different error costs for different classes, and adjusting the class boundary based on kernel-alignment ideal were proposed.

2.3 Feature Selection for Imbalance Datasets

The purpose of feature selection is to avoid over-fitting and improve model's performance, to provide a cost-effective model, and to gain a deeper insight into the underlying processes that generated the data. In the field of imbalanced datasets mining, feature selection is even more important than the choice of the learning method. The methods are divided into three groups: filter methods, wrapper methods, and embed methods.

2.4 Evaluation Metrics

When performing the classification on the imbalanced dataset, overall accuracy is no longer suitable for evaluating the performance of classifier. Thus, besides overall accuracy, in this study, G-mean and Matthews correlation coefficient are used, which are defined as follows:

$$\text{G-mean (Balanced accuracy)} = \left(\frac{TP \times TN}{(TP+FN)\times(TN+FP)}\right)^{1/2}$$

$$\text{Matthews Correlation Coefficient } (MCC) = \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\left((\text{TP}+\text{FN})\times(\text{TP}+\text{FP})\times(\text{TN}+\text{FP})\times(\text{TN}+\text{FN})\right)^{1/2}}$$

where TP is the number of positive samples that are correctly predicted as positive; TN is the number of negative samples that are correctly predicted as negative; FP is the number of negative samples that are predicted as positive; and FN is the number of positive samples that are predicted as negative.

In addition, the threshold independent measures ROC (Receiver Operating Characteristics) curve and AUC (Area Under the Curve) [10], are adopted. An acceptable classification model should have AUC above 0.5. An AUC value above 0.7 indicates a useful prediction, and a good prediction method achieves AUC above 0.85.

Chapter 3 Improving the Prediction of Protein-Protein Interaction Sites Using a Novel Over-sampling Approach and Predicted Shape Strings

3.1 Introduction

A naive approach based on support vector machines often fails to predict binding interfaces among interacting proteins with high specificity since the number of non-interaction residues is much larger than the number of interaction residues. In this study, we propose a novel over-sampling approach in order to relax class-imbalance for the dataset of PPI sites. Instead of dealing with all minority class samples equivalently, we intentionally increase the number of minority samples according to their local distribution. Furthermore, predicted shape strings are used to enrich the feature groups.

3.2 Materials and Methods

3.2.1 Dataset

We used two datasets that were highly imbalanced for evaluating the performance. The first one (that was named D1050) was the same with Chen and Jeong [11]. Each residue was represented as a 1,050 features. The dataset contained 2,829 interface residues (positive class) and 24,616 non-interface residues (negative class).

The second dataset (was named D1239) was prepared by adding information of predicted shape strings to the original dataset. DSP program [12] was used to predict the shape strings. Each sample in this dataset includes 1, 239 features.

3.2.2 Methods

Resampling techniques

In order to alleviate the problem of overlapping and over-fitting simultaneously, we propose a novel over-sampling algorithm, which we call Over-sampling based on local Density (OSD). OSD algorithm focuses on only minority samples located where the local density of minority samples is small in comparison with that of majority samples.

Definition 1. Suppose m and n are the numbers of samples with the same and different class labels for sample x, respectively. Local density of x with radius r is the proportion m/(m+n).

OSD- a novel over-sampling approach

A key idea of the OSD algorithm is to increase the number of minority samples located where the local density of minority samples is small in comparison with majority samples. The number of synthetic samples for each x depends on its local distribution with parameter d:

- If x doesn't have neighbor (i.e. $m + n = 0$), or local density of x is 0 (i.e. $m = 0$), x locates far from the other minority samples and OSD generates the maximum number of synthetic samples with the same class labels as x in order to avoid the class imbalance problem and diminish boundary variance derived from local sparsity, simultaneously. Hence, d new samples will be synthesized.
- If local density of x is greater than 0, $d*(1-m/(m+n))$ new synthetic samples are created.
- If sample x has no different class label neighbor, OSD does not adjust the local density of x.

KSVM-THR

We note that OSD generally does not balance imbalanced datasets entirely. To address this issue, we combine OSD and KSVM-THR, SVM with adjustment of the decision parameter [13]. The decision threshold θ of KSVM-THR is defined as

$$\theta = -1 + 2 * (p + \alpha)/(p + n + 2 * \alpha)$$

where p and n are the numbers of minority and majority class samples, respectively. The constant α is the tuning parameter and in the experiments below, it was optimized by grid search.

Experimental design

SVM with Gaussian RBF kernel was utilized to create a basic classifier. We conducted 10-fold cross validation. All the features of the datasets were normalized. Noise samples in the datasets were filtered out before over-sampling, where we defined samples that have the same feature vector and that belongs to different classes as noise samples. The overall predicting process is shown in Figure 3.1.

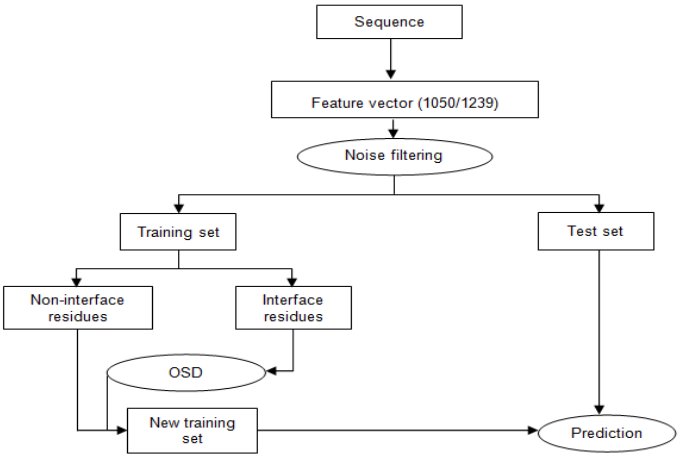

Figure 3.1 Schematic representation of our method.

3.3 Results and Discussions

3.3.1 Evaluation on the D1050 Dataset

Experimental results on the dataset D1050 were compared with KSVM without resampling (KSVM-only), Random Under-sampling (RUS), KSVM-THR-only, weighted SVM, SMOTE, the method of Chen and Jeong, and the under-sampling method introduced by Anand et al. [9]. The results of all these methods are shown in Table 3.1.

Figure 3.2 demonstrates the ROC curves of OSD and the other methods. ROC curve of Cheng and Jeong was taken from [11].

Table 3.1 Performance measures comparison of different methods on the dataset D1050 in terms of best G-mean

Method	Overall accuracy (%)	Sensitivity (%)	Specificity (%)	G-mean
KSVM-only	90.11	4.66	99.93	21.59
OSD	88.23	67.86	90.57	78.40
RUS (1.1:1)	76.17	70.59	76.81	73.63
RUS-OSD	75.31	80.73	74.69	77.65
KSVM-THR-only	90.66	11.48	99.76	33.85
OSD-THR	83.36	77.73	84.01	80.80
RUS-THR(1.1:1)	65.71	82.11	83.82	72.39
RUS-OSD-THR	64.94	88.51	62.24	74.22
Weighted-SVM*	91.57	55.87	95.56	73.08
SMOTE*	92.96	51.74	97.69	71.07
Chen and Jeong (2009)*	71.90	71.20	71.98	71.59
Anand et al. (2010)*	77.53	71.04	78.27	74.54

**: Result was taken from the paper of Anand et al.*

3.3.2 Evaluation on the D1239 Dataset

Table 3.2 shows the improvements using our algorithm and new decision threshold in the comparison of the naïve classifier. It indicates that our over-sampling algorithm based on the local density can relieve the class-imbalance problem in this dataset.

Table 3.3 displays the comparative results on the datasets D1050 and D1239. In Figure 3.3, it can be seen that the performance of KSMV-only on D1239 is apparently better than the one on D1050 in the area of recall lower than 0.3 and precision higher than 0.8. It means that shape string is effective for performance improvement in this area.

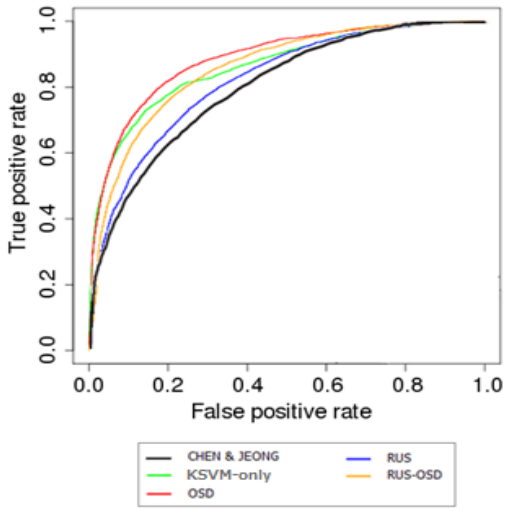

Figure 3.2 ROC curves of the competing methods on the D1050 dataset

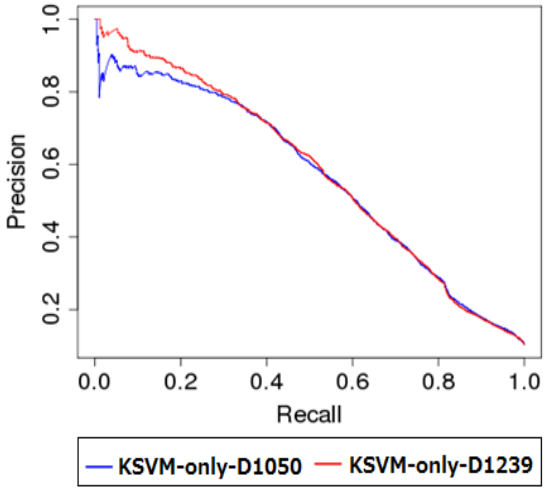

Figure 3.3 PR curves for the datasets with shape string (D1239) and without shape string (D1050) prediction with KSVM as basic classifier

Table 3.2 Performance measures comparison of different methods on the dataset D1239

Method	Overall accuracy (%)	Sensitivity (%)	Specificity (%)	G-mean
KSVM-only	90.45	8.02	**99.92**	28.31
OSD	89.61	63.13	92.66	76.48
KSVM-THR-only	**91.07**	15.30	99.78	34.79
OSD-THR	85.49	**74.44**	86.76	**80.36**

Table 3.3 Performance measures comparison on the datasets D1239 and D1050

Data set	Method	Precision (%)	Recall (%)	F-measure (%)
D1050	KSVM-only	**89.18**	4.66	8.86
	OSD	45.27	67.86	**54.31**
	KSVM-THR-only	85.07	11.48	20.24
	OSD-THR	35.84	**77.73**	49.06
D1239	KSVM-only	**92.65**	8.02	14.76
	OSD	49.72	63.13	**55.63**
	KSVM-THR-only	89.09	15.30	26.12
	OSD-THR	39.26	**74.44**	51.40

3.4 Conclusion

In this study, we aimed at the identification of protein-protein interaction sites. The PPI datasets used in this study were highly class-imbalanced, which often decrease classification performance of SVMs. To avoid this issue, we proposed a novel over-sampling technique that effectively utilizes local density of minority samples. We also proposed several methods combined with KSVM-THR and random under-sampling methods to reinforce the tolerance for the class imbalance problem. Experimental results showed that the combination of our OSD algorithm and new feature group led to higher sensitivity, G-mean, precision, MCC, F-measure, and AUC-PR, at least comparable performance with the state-of-the-art methods. In addition, we found that the information of predicted shape strings increase the performance for predicting whether interface or non-interface residues. Further extensions can be considered, for example, combining our algorithm with other heuristic under-sampling method, or feature selection methods.

Chapter 4 Improvement in β-turns Prediction Using Predicted Protein Blocks and Random Under-sampling Method

4.1 Introduction

In this study, we introduce a novel method that can enhance the result of predicting β-turns and their types by using the informative feature groups and dealing with class imbalance problem where the ratio of non-turn residues to the turn residues and the non-specific-type-turn residues to the correct-type-turn residues are high. We present the experimental results on three standard benchmark datasets in comparison with state-of-the-art methods.

4.2 Materials and Methods

4.2.1 Datasets

We utilized three standard benchmark datasets BT426, BT547 and BT823 to evaluate the performance of our method. The numbers of protein sequences in these datasets are 426, 547 and 823, respectively. In this study, type VI was not considered because of its rare appearance.

4.2.2 Feature vector

Position Specific Scoring Matrices (PSSMs)

The PSSMs were generated by using PSI-BLAST with default parameters. Each element x of these matrices was scaled within the range [0,1] by the logistic function:

$$f(x) = \frac{1}{1 + e^{-x}}$$

Predicted Protein Blocks

PB-kPRED [14] was used to get the predicted protein blocks. Sixteen characters from A to P symbolized sixteen blocks and X represented the unidentified state. For each residue i in a protein chain, its predicted protein block was represented by a vector of 17 features $(x_i^j)_{17}$, where x_i^j was the probability of residue i as state j.

The feature vector corresponds to each query residue was generated by using a sliding window of size nine amino acids. Thus, there were 333 attributes in one input vector.

4.2.3 Experimental design

We conducted seven-fold cross validation to evaluate the performance of our method. Each dataset was divided into seven parts that contained the same number of positive samples. Support Vector Machines with Gaussian RBF kernel were employed as the basic classifier.

Random Under-Sampling (RUS) was utilized to balance the training datasets before predicting. Grid search relying on MCC to choose the optimal ratio for RUS was operated. Then, feature selection based on information gain ratio was applied to reduce the redundant features and achieve the highest MCC.

Figure 4.1 demonstrates the overall architecture of our method.

4.2.4 Filtering

These following rules were applied to ensure that every final predicted turn is longer than four residues:

i. Change isolated non-turn prediction to turn: tnt → ttt

ii. Change isolated turn prediction to non-turn: ntn → nnn

iii. Change the two non-turn neighbors of two successive turns to turns: nttn → tttt

iv. Change the two non-turn neighbors of three successive turns to turns: ntttn → ttttt

4.2.5 Performance metrics

As MCC, which is said to be the most robust, Q_{total} (overall accuracy), Q_{obs} (sensitivity), Q_{pred} (precision) are often used to measure the quality of β-turn prediction methods, they are also used to evaluate the performance of our method.

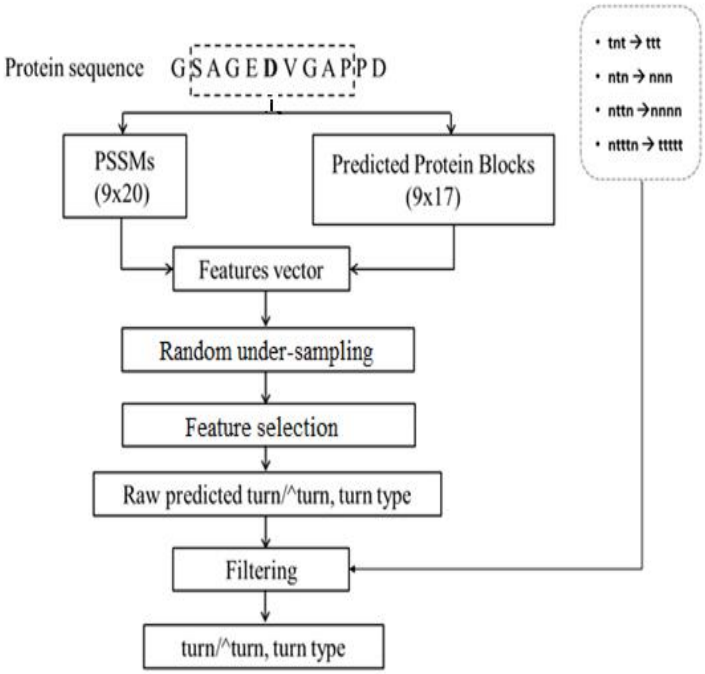

Figure 4.1 General scheme of our method

4.3 Results and Discussions

4.3.1 Turn/non-turn prediction

We employed experiments with various sliding window sizes and selected the size of nine residues since it returns the highest results.

Experiments to value the impact of evolutionary information PSSMs, predicted protein block, and their combination were also performed.

The comparison of our method with the other competitive methods on the BT426 dataset is presented in Table 4.1. It shows that our method outperformed KLR and the others with MCC of 0.585.

Table 4.1 Comparison of competitive methods on the BT426 dataset. "_" means this value was not reported

Method	Q_{total} (%)	Q_{obs} (%)	Q_{pred} (%)	Specificity (%)	MCC	AUC
Our method	**84.41**	71.01	**66.89**	**88.71**	**0.585**	**0.893**
KLR [15]	80.4	65.25	58.98	85.34	0.50	0.86
NetTurnP [16]	78.2	**75.6**	54.4	79.1	0.50	0.86
DEBT [17]	79.2	70.1	54.8	-	0.48	0.84
BTNpred [18]	80.9	55.6	62.7	-	0.47	-
SVM [19]	79.8	68.9	55.6	-	0.47	0.87
BTSVM [20]	78.7	62.0	56.0	-	0.45	-
BetaTPred [3]	75.5	72.3	49.8	-	0.43	-
BTPRED [21]	74.9	48.0	55.3	-	0.35	-

Table 4.2 presents the results of the competing methods on the datasets BT547 and BT823, with our method achieved the highest values on MCC, Q_{total} and Q_{pred}.

Table 4.2 Comparison of competitive methods on the BT547 and BT823 datasets. "_" means this value was not reported

Dataset	Method	Q_{total} (%)	Q_{obs} (%)	Q_{pred} (%)	Specificity (%)	MCC	AUC
BT547	Our method	**85.01**	64.70	**73.37**	**91.96**	**0.591**	**0.894**
	KLR [15]	80.46	65.36	59.04	-	0.50	-
	DEBT [17]	80.0	68.7	55.9	-	0.49	0.85
	BTNpred [18]	80.5	54.2	61.6	-	0.45	-
	SVM [19]	76.6	70.2	47.6	-	0.43	-
	COUDES [22]	74.6	**70.4**	48.7	-	0.42	-
BT823	Our method	**84.96**	68.46	**70.51**	**90.46**	**0.595**	**0.896**
	KLR [15]	80.66	64.64	58.42	-	0.49	-
	DEBT [17]	80.9	66.1	55.9	-	0.48	0.84
	BTNpred [18]	80.6	54.6	60.8	-	0.45	-
	SVM [19]	76.8	**72.3**	53.0	-	0.45	-
	COUDES [22]	74.2	69.6	47.5	-	0.41	-

4.3.2 Turn types prediction

Table 4.3 presents the MCC of competing methods on the three datasets BT426, BT547, BT823.

Table 4.3 MCCs comparison between the competitive methods

Dataset	Method	I	I'	II	II'	IV	VIII
BT426	Our method	0.551	**0.635**	0.561	**0.530**	0.315	0.223
	X.Shi et al. [23]	**0.714**	0.513	**0.684**	0.415	**0.459**	**0.246**
	NetTurnP[16]	0.36	0.23	0.31	0.16	0.27	0.16
	DEBT[17]	0.36	_	0.29	_	0.27	0.14
	COUDES [22]	0.309	0.226	0.302	0.106	0.109	0.071
BT547	Our method	**0.545**	**0.632**	**0.578**	**0.453**	**0.322**	**0.235**
	X.Shi et al. [23]	0.529	0.538	0.548	0.337	0.311	0.044
	DEBT[17]	0.38	_	0.33	_	0.27	0.14
BT823	Our method	0.554	**0.635**	0.587	**0.454**	**0.326**	**0.225**
	X.Shi et al. [23]	**0.636**	0.416	**0.630**	0.361	0.317	0.125
	DEBT[17]	0.39	_	0.33	_	0.27	0.14

4.4 Conclusions

In this study, we presented a new method to identify the β-turns and their types in protein sequence. We focused on both using more the well-characterized features and class-imbalanced-dealt technique. We achieved the highest MCCs of 0.585, 0.591 and 0.595 on the three datasets BT426, BT547 and BT823, respectively, in comparison with the state-of-the-art β-turns prediction methods. In the field of β-turn types prediction, we also harvested the high and stable results. Further extension can be considered such as using the effective method to handle the class-imbalanced problem.

Chapter 5 Conclusions

5.1 Dissertation Summary

Proteins are very important because they are involved in many functions in a living cell. Most proteins do not work alone but in the collaboration with the other ones. However, many interactions between proteins are unidentified until now. Therefore, study the mechanism of protein-protein interactions, especially, which part in protein sequence has the contacted ability, is one of the necessary problems in bioinformatics.

Nevertheless, to deeply understand the protein-protein interaction sites as well as the other functions of proteins, it is necessary to understand their three-dimensional structure. One of the most important tasks in this field is learning about β-turns and their types.

In this thesis, we aimed at (i) improving the performance of protein-protein interaction sites prediction using a novel over-sampling method and informative features; and (ii) improving the prediction of the β-turns and their types by applying predicted protein blocks and under-sampling technique.

5.2 Future Works

In this thesis, we developed the new algorithm OSD to over-sample the minority set of an imbalanced dataset by focusing on the local density. However, OSD just can handle the numerical values but the nominal values. Thus, the extension of OSD can be thought about so that it can be applied for the datasets with nominal features. In addition, because feature selection affects the performance of prediction on imbalanced dataset, we can combine feature selection with our methods, as a preprocessing step. It may lead to improve the results. In addition, random under-sampling is the most naïve under-sampling method. This method is simple and fast, however, lost a lot of information. Thus, the use of better under-sampling method may result in better performance than random under-sampling.

About the second problem in our thesis, the β-turns prediction, we also think about applying the under-sampling technique that is better than random under-sampling. Since the model that was created by utilizing PSSMs, predicted protein block, under-sampling and feature selection returns good results in this situation, it also can be used for predicting protein-protein interaction sites and the other kind of tight turns such as α-turns or γ-turns.

In addition, residues belong to β-turn type VI were not predicted because of the limitation of their appearances in a protein chain. Thus, we aim to develop our method that in the future, we can recognize all the β-turn types.

Bibliography

1. De Brevern AG, Etchebest C, Hazout S: **Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks.** *Proteins* 2000, **41**:271–87.

2. De Brevern AG: **New Assessment of a Structural Alphabet**. *In Silico Biology* 2005, **5**:283–289.

3. Kaur H, Raghava GPS: **Prediction of beta-turns in proteins from multiple alignment using neural network**. *Protein Science* 2003, **12**:627–634.

4. Hutchinson EG, Thornton JM: **A revised set of potentials for beta-turn formation in proteins**. *Protein Science* 1994, **3**:2207–2216.

5. Vapnik V, Lerner A: **Pattern Recognition using Generalized Portrait Method**. *Automation and Remote Control* 1963, **24**.

6. Chawla N V., Bowyer KW, Hall LO, Kegelmeyer WP: **SMOTE : Synthetic Minority Over-sampling Technique**. *Journal of Artificial Intelligence Research* 2002, **16**:321–357.

7. Liu X, Wu J, Zhou Z: **Exploratory Undersampling for Class-Imbalance Learning**. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 2009, **39**:539–550.

8. Mani I, Zhang J: **kNN approach to unbalanced data distributions: a case study involving information extraction**. In *Proceedings of Workshop on Learning from Imbalanced Datasets*. 2003.

9. Anand A, Pugalenthi G, Fogel GB, Suganthan PN: **An approach for classification of highly imbalanced data using weighting and undersampling**. *Amino Acids* 2010, **39**:1385–1391.

10. Sonego P, Kocsor A, Pongor S: **ROC analysis: applications to the classification of biological sequences and 3D structures**. *Briefings in Bioinformatics* 2008, **9**:198–209.

11. Chen X, Jeong JC: **Sequence-based prediction of protein interaction sites with an integrative method.** *Bioinformatics (Oxford, England)* 2009, **25**:585–91.

12. Sun J, Tang S, Xiong W, Cong P, Li T: **DSP: a protein shape string and its profile prediction server**. *Nucleic Acids Research* 2012, **40**:W298–W302.

13. Lin W-J, Chen JJ: **Class-imbalanced classifiers for high-dimensional data**. *Briefings in Bioinformatics* 2013, **14**:13–26.

14. **PB-PENTAPEPT** [http://www.bo-protscience.fr/pentapept/?page_id=9].

15. Elbashir MK, Wang J, Wu F, Li M: **Sparse Kernel Logistic Regression for β -turns Prediction**. *Systems Biology (ISB), 2012 IEEE 6th International Conference on* 2012:246–251.

16. Petersen B, Lundegaard C, Petersen TN: **NetTurnP – Neural Network Prediction of Beta-turns by Use of Evolutionary Information and Predicted Protein Sequence Features**. *PloS ONE* 2010, **5**:e15079.

17. Kountouris P, Hirst JD: **Predicting beta-turns and their types using predicted backbone dihedral angles and secondary structures**. *BMC Bioinformatics* 2010, **11**:407.

18. Zheng C, Kurgan L: **Prediction of beta-turns at over 80% accuracy based on an ensemble of predicted secondary structures and multiple alignments**. *BMC Bioinformatics* 2008, **9**:430.

19. Hu X, Li Q: **Using support vector machine to predict beta- and gamma-turns in proteins**. *Journal of Computational Chemistry* 2008, **29**:1867–75.

20. Pham TH, Satou K, Ho TB: **Prediction and analysis of beta-turns in proteins by support vector machine**. *Genome Informatics* 2003, **14**:196–205.

21. Shepherd AJ, Gorse D, Thornton JM: **Prediction of the location and type of beta-turns in proteins using neural networks**. *Protein Science* 1999, **8**:1045–1055.

22. Fuchs PFJ, Alix AJP: **High accuracy prediction of beta-turns and their types using propensities and multiple alignments.** *Proteins* 2005, **59**:828–39.

23. Shi X, Hu X, Li S, Liu X: **Prediction of β-turn types in protein by using composite vector**. *Journal of Theoretical Biology* 2011, **286**:24–30.

学位論文審査結果の要旨

平成２５年７月３０日に第１回学位論文審査委員会を開催、８月６日に口頭発表、その後に第２回審査委員会を開催し、慎重審議の結果、以下の通り判定した。なお、口頭発表における質疑を最終試験に代えるものとした。

機械学習手法を用いた分類において精度が下がる重要な要因の１つに、クラス不均衡があり、生物学データや医学データを対象とする場合にもこの問題は頻繁に起こる。本研究では、タンパク質間相互作用予測問題と、タンパク質の立体構造におけるβターン領域予測問題を対象として、クラス不均衡を緩和することにより予測精度の向上を図った。前者に対しては局所的なサンプル密度を導入した新しいオーバーサンプリング手法を提案することにより、従来法を上回る予測精度を達成した。一方、後者に対してはアンダーサンプリングと特徴選択を組み合わせた新しい手法を提案することにより、同様に従来法を上回る予測精度を達成した。

以上の研究成果は、生命医科学分野を含む広い範囲における高精度なデータ分類と解析に大きく貢献するものであり、本論文は博士（工学）に値するものと判定した。